Statistical tests for Rare Variants Data Rare Variants in Human Genetic Diseases: Comparison of Association Statistical Tests

Lise Bellanger, Elodie Persyn, Floriane Simonet, Richard Redon, Jean-Jacques Schott, Solena Le Scouarnec, Matilde Karakachoff, Christian Dina

Rare Variants in Human Genetic Diseases: Comparison of Association Statistical Tests

Lise BELLANGER[1], Elodie PERSYN[2], Floriane SIMONET[2], Richard REDON[2] , Jean-Jacques SCHOTT[2], Solena LE SCOUARNEC[2], Matilde KARAKACHOFF[2], Christian DINA[2]

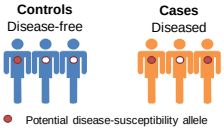
[1] Laboratoire de Mathématiques Jean Leray (LMJ) UMR CNRS 6629, University of Nantes, France. E-mail: lise.bellanger@univ-nantes.fr

[2] Institut du thorax, Inserm UMR 1087 / CNRS UMR 6291, University of Nantes, France

Background and Objectives

Genome-wide association studies (GWAS) aim to identify genetic loci of susceptibility of complex diseases. These studies require:

- sampling of individuals from 2 types of individuals: **cases** and **controls**



- genotyping of individuals for a dense set of genetics variants (Single Nucleotide Variants (**SNVs**)) across the whole genome.

GWASs allow to identify common genetic variants (CVs) associated with many common diseases. But these variants do not fully explain the heritability (*i.e.* proportion of phenotypic variance attributable to genetic variance).

GWAS	Missing heritability	***Exome sequencing studies***
CDCV hypothesis (Common Disease Common Variant)	⇒	**CDRV** hypothesis (Common Disease Rare Variant)

Exome sequencing studies: Exons of genes are sequenced for cases and controls.

Minor Allele Frequency (MAF): frequency at which the least common allele occurs in a given population.

Rare variant (RV): mutation present in less than 1-5% of the population. (SNV with MAF≤ 1-5%).

Problems:

- Most popular statistical tests for GWAS based on testing single SNP are powerless due to **rare + low effects**
- Development of many statistical tests specifically targeting RVs to detect association between a set of variants (*e.g.* located on a gene) and a disease.

Our Objectives:

Which ? Comparative evaluation of the existing tests for the identification of disease-associated RVs in order to identify the most suitable ones in the context of associated RVs.

Why ? There exists a lot of tests but none is uniformly most powerful. Adaptive tests are needed!

How ? With simulated data and real data.

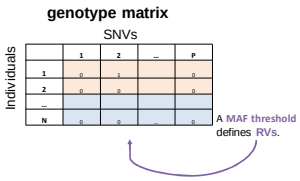
Material and Methods

Statistical tests for Rare Variants

(*) Also named *burden tests*.

Notations

$i \in \{1, \dots, N\}$ individuals
$j \in \{1, \dots, P\}$ SNVs
X= genotype matrix (N x P) for **one gene**
$X_{ij} \in \{0,1,2\}$: genotype of individual i for the SNV j
$Y_i \in \{0,1\}$: phenotype of subject i



Test	Pool	MAF threshold	Sensitive to +/-	weights	Permut	Refs
CAST	yes	fixed	no	no	no	1
CMC	yes	fixed	no	no	poss.(*)	2
WSS	yes	no	no	yes	yes	3
aSUM	yes	no	yes	poss.	yes	4
VT	yes	variable	no	no	yes	5
KBAC	no	fixed	no	yes	yes	6
C-Alpha	no	fixed	yes	poss.	yes	7
SKAT	no	no	yes	poss.	poss.	8
SKAT-O	no	no	yes	poss.	poss.	9
BOMP	hybrid	no	yes	poss.	yes	10

(*) poss.=possible, but not necessarily by default.

Summary of properties of the tests for RVs to be compared:

Whether pooling over variants, using a MAF threshold to define RVs, sensitive to association directions (+/-), whether possible use of weights, requiring permutations for *p*-value calculations and references for more details.

Pooled association tests (*)

WSS
Weighted-Sum Statistic is analogous to the Mann-Whitney-Wilcoxon rank test, with weights on RVs.

CAST
Cohort Allelic Sum Test compares the proportion of cases that present at least one RV with that of controls.

CMC
Combined Multivariate and Collapsing test modifies the *CAST* to improve its performance when both RVs and CVs are present.

VT
Variable Threshold method is based on the intuition that there exists an (unknown) threshold T for which variants with MAF ≤ T are substantially more likely to be functional.

aSUM
Adaptive Sum test introduces data-adaptive weights to differentiate between beneficial, neutral or deleterious variants.

BOMP
Burden Or Mutation Position test is an hybrid likelihood approach that combines a burden test with a test of the position distribution of variants.

C-Alpha test
C-alpha statistics tests the presence of a mixture of risk, protective or neutral variants.

KBAC
Kernel-based adaptive Clustering method groups mutation patterns across the variants and assigns each mutation pattern a kernel-based weight adaptively determined by data.

SKAT SKAT-O
Generalization of *C-alpha test*, *Sequence Kernel Association Test* uses a multiple regression approach with introduction of kernel. The optimal test (*SKAT-O*) is a generalized family of *SKAT* tests that incorporates a correlation structure of variants effects.

Data

Simulated data

Genotypes and phenotypes generated as in (11), **6 main scenarios**, sub-divided into several scenarios, 500 independent replicates, test significant level $\alpha = 5\%$, 500 permutations for permutation-based methods:

- sample size 500 cases and 500 controls;
- 8 causal RVs, p∈ {0,4,8,16,32} non causal RVs;
- $\rho = 0$ independent or $\rho = 0.9$ in Linkage Disequilibrium (LD).

Variant category	RV	LFV (Low Frequency Variant)	CV (Common Variant)
MAF	[0,001; 0,01]	]0,01; 0,05]	]0,05; 0,1]

***Scenario 1*: Independent RVs:** no LD between RVs.
***Scenario 2*: RVs in LD :** all RVs (both causal and non-causal) correlated.
***Scenario 3*: No LD between causal RVs and non-causal RVs:** causal RVs correlated, non-causal RVs correlated.
***Scenario 4*: Independent RVs and LFVs (non-causal):** independent RVs and 4 non-causal LFVs.
***Scenario 5*: Independent RVs and LFVs:** independent RVs and 4 LFVs (3 non-causal and 1 causal).
***Scenario 6*: Independent RVs, LFVs and CVs:** independent RVs, 4 non-causal LFVs and 2 non-causal CVs.

Evaluation:
→ Type I error rate analysis (OR=1 in the logistic regression model generating phenotypes: null case);
→ Empirical power analysis (OR≠1: non-null cases).

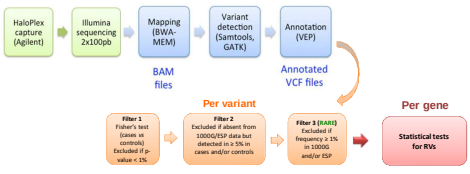
Real sequencing data

Cardiac arrhythmias

Population sampling design
167 **Controls** Calcific Aortic Valve Stenosis patients (CAVS)
167 **Cases** Brugada syndrome patients (BrS)

Targeted sequencing design
163 targeted genes
Brugada syndrome: a rare heritable arrhythmia syndrome associated with an increased risk of sudden cardiac death (12).

Pipeline of data creation and analysis



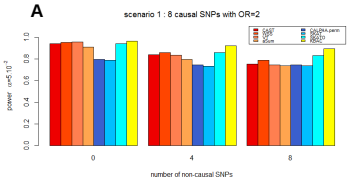
After preprocessing, 135 matrices (one per studied gene)
1 to 280 variants in column and the 334 patients in line.

Results

Scenario 1: Satisfactory Type I error rates for all tests.

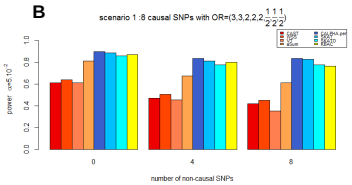
Correlation PCA on $-\log_{10}$(*p*-values) matrix
135 genes x tests for RVs - MAF threshold defining RVs = 0,01

Power analysis



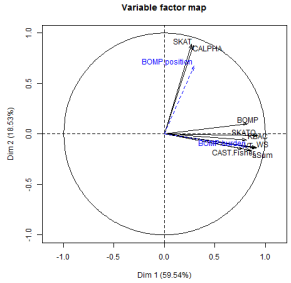
A: 8 deleterious causal RVs with common OR=2.

- Ideal situation for pooled association tests. Their performance deteriorates when the number of non-causal RVs increases.
- *KBAC* is the most powerful.



B: 8 causal RVs with no constant association strengths.

- Pooled association tests *CAST*, *WSS* and *VT* suffer from loss of power.
- *C-Alpha*, *SKAT*, *SKAT-O* and *KBAC* perform similarly and are most powerful.



- The first dimension remains linked to
  - all studied burden tests and
  - some non-burden tests: *SKAT-O*, *KBAC* and *BOMP*.
- The second dimension appears as the dimension of two non-burden tests:
  - *C-Alpha* and its generalization *SKAT*.

Conclusion and Perspectives

- *Simulated data:* simulations are time-consuming even with the use of parallel computing!
  Perspectives: tests comparison for all scenarios and also for a population genetics model.
- *Real sequencing data:* detection of differences between tests in a practical set-up. It reveals difficulties linked to preprocessing steps and also to the experimental design (choice of controls).
  Perspectives: recommendations on the use of RVs tests and strategy to detect RVs on real data.
- The definition of a more powerful test depends on the unknown truth (of the association pattern): improvements to existing tests like adaptive tests are needed!

References

1. Morgenthaler and Thilly (2007) *Mutat. Res.* **615**: 28–56.
2. Li and Leal (2008) *Am. J. Hum. Genet.* ***83***: 311–321.
3. Madsen and Browning (2009). *PLoS Genet.* **5(2)**: e1000384
4. Han and Pan (2010) *Hum. Hered.* **70**: 42-54.
5. Price *et al.* (2010) *Am. J. Hum. Genet.* **86**: 832–838.
6. Chen *et al.* (2013) *PLoS Genet.* **9**, *e1003224*.
7. Liu and Leal (2010) *PLoS Genet.* **6**, *e1001156*.
8. Neale *et al.* (2011) *PLoS Genet.* **7**, *e1001322*.
9. Wu *et al.* (2011) *Am. J. Hum. Genet.* **89**: 82–93.
10. Lee *et al.* (2012) *Am. J. Hum. Genet.* **91**: 224–237.
11. Basu and Pan (2011) *Genet. Epidemiol.* **35**: 606–619.
12. Crotti *et al.* (2012) *JACC.* **60**(15): 1410-1418.